CENSUS IN COUNTIES — DESCRIBING AND COMPARING HISTOGRAMS TO UNDERSTAND AMERICAN LIFE

TEACHER VERSION

Subject Level:
High School Math

Grade Level:
9–12

Approx. Time Required:
45–60 minutes

Learning Objectives:

- Students will be able to analyze data and describe data distributions.
- Students will be able to understand skewness and how it affects the mean and median.
- Students will be able to determine the direction of skewness from a distribution's mean and median.
- Students will be able to understand the impact of outliers on the standard deviation and interquartile range.

Activity Description

Students will analyze a variety of county-level census data, including on employment, technology, and transportation, in histograms to compare and contrast the shapes of their distributions and to interpret measures of center and spread in context.

Suggested Grade Level:
9–12

Approximate Time Required:
45–60 minutes

Learning Objectives:

- Students will be able to analyze data and describe data distributions.
- Students will be able to understand skewness and how it affects the mean and median.
- Students will be able to determine the direction of skewness from a distribution's mean and median.
- Students will be able to understand the impact of outliers on the standard deviation and interquartile range.

Topics:

- Histograms
- Measures of center
- Measures of spread
- Shape of distributions
- Skewness

Skills Taught:

- Interpreting skewness from a graph
- Interpreting the standard deviation and interquartile range
- Understanding skewness and its effect on the mean and the median
- Understanding when to use the mean versus the median

Materials Required

- The student version of this activity, 16 pages

Activity Items

There are no separate items for this activity.

For more information to help you introduce your students to the U.S. Census Bureau, read "*Census Bureau 101 for Students*." This information sheet can be printed and passed out to your students as well.

Standards Addressed

See chart below. For more information, read "*Overview of Education Standards and Guidelines Addressed in Statistics in Schools Activities*."

Common Core State Standards for Mathematics

Standard	Domain	Cluster
CCSS.MATH.CONTENT.HSS.ID.A.2 Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, standard deviation) of two or more different data sets.	ID – Interpreting Categorical & Quantitative Data	Summarize, represent, and interpret data on a single count or measurement variable.
CCSS.MATH.CONTENT.HSS.ID.A.3 Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).	ID – Interpreting Categorical & Quantitative Data	Summarize, represent, and interpret data on a single count or measurement variable.

Common Core State Standards for Mathematical Practice

Standard
CCSS.MATH.PRACTICE.MP2. Reason abstractly and quantitatively. Students will use quantitative reasoning to determine the relationship between the mean and median of a data set based on its histogram.

National Council of Teachers of Mathematics' Principles and Standards for School Mathematics

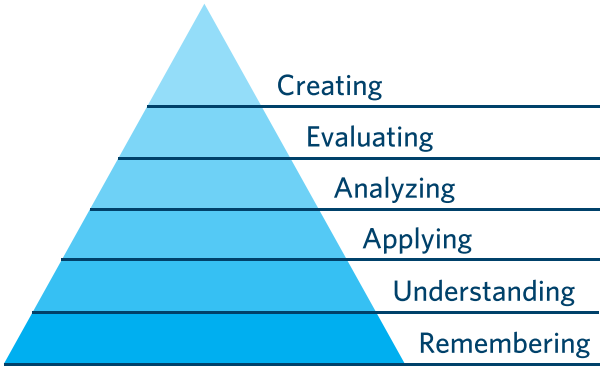
Content Standard	Students should be able to:	Expectation for Grade Band
Data Analysis and Probability	Select and use appropriate statistical methods to analyze data.	For univariate measurement data, be able to display the distribution, describe its shape, and select and calculate summary statistics.

Guidelines for Assessment and Instruction in Statistics Education

GAISE	Level A	Level B	Level C
Formulate Questions			
Collect Data			
Analyze Data	X	X	
Interpret Results			

Bloom's Taxonomy

Students will ***understand*** how the mean and median relate in skewed distributions. They will also ***analyze*** and ***evaluate*** census data to draw conclusions about which measure of center is more representative for various distributions.



Teacher Notes

Before the Activity

Students must understand the following key terms:

- **Histogram** – a way of displaying numerical data on a graph using horizontal or vertical bars of equal width so that the height or the length of the bars indicates frequency
- **Center** – the middle of a data set, usually identified as the median or mean
- **Mean** – a measure of center in a set of numerical data, computed by adding the values in a list and then dividing by the number of values in the list
- **Median**– a measure of center in a set of numerical data, identified as the value appearing at the middle of a sorted version of the list (or the mean of the two middle values if the list contains an even number of values)
- **Shape** – the general form of a data distribution (e.g., bell-shaped, bimodal, irregular, uniform)
- **Measure of variability** – a measure of how spread out the values are in a data set
- **Spread** – the amount of variability among the values in a data set
- **Standard deviation (SD)** – a measure of spread for a set of numerical data, calculated by taking the square root of the variance, that increases in value as the data in the set become more spread out
- **Range** – the numerical difference between the lowest and highest values in a distribution
- **Interquartile range (IQR)** – a measure of variability in a set of numerical data to indicate the difference between the lower and upper quartiles of the data set
- **5-number summary** – a summary that provides information about a data set and contains five values: the minimum (smallest value), first quartile, median (middle value), third quartile, and maximum (largest value)
- **Skewness** – a measure of the unevenness of values in a data set in which one "tail" of the distribution has more extreme values than the other "tail," described using terms like "skewed right" and "skewed left"
- **Symmetry** – when the data points on both sides of a distribution appear approximately the same
- **Outlier** – a data point that is well outside of the expected range of values or does not follow the overall pattern of the other data points
- **Family household** – a household in which at least one person is related to the householder by birth, marriage, or adoption
- **Own children** – children of a household who are younger than 18 and who have never been married; this designation represents a household type on the census questionnaire.
- **Variance** – a measure of variability in a data set, calculated by first finding the numerical difference between each value and the mean, squaring the differences, and then finding the average of all the squared differences

Students should have a basic understanding of the following ideas and concepts:

- How outliers affect the SD, range, and IQR
- How skewness affects the mean and median

Data in this activity come from the 2014 American Community Survey, which is conducted monthly by the U.S. Census Bureau and is designed to show how communities are changing. Through asking questions of a sample of the population, it produces national data on more than 35 categories of information, such as education, income, housing, and employment.

Note: To access the data for a county or multiple counties click on the link under each graph, select "Geographies" at the top left of the screen and then select "County" and choose the state(s).

During the Activity

Teachers should monitor students as they work.

After the Activity

Teachers should ask students to reflect on what they learned.

Extension Idea

- Teachers could have students experiment with this page on the NCTM Illuminations web site (*illuminations.nctm.org/Activity.aspx?id=3576*), which illustrates the influence of different distribution types on the mean and the median.

Student Activity

Click ***here*** to download a printable version for students.

Activity Items

There are no separate items for this activity.

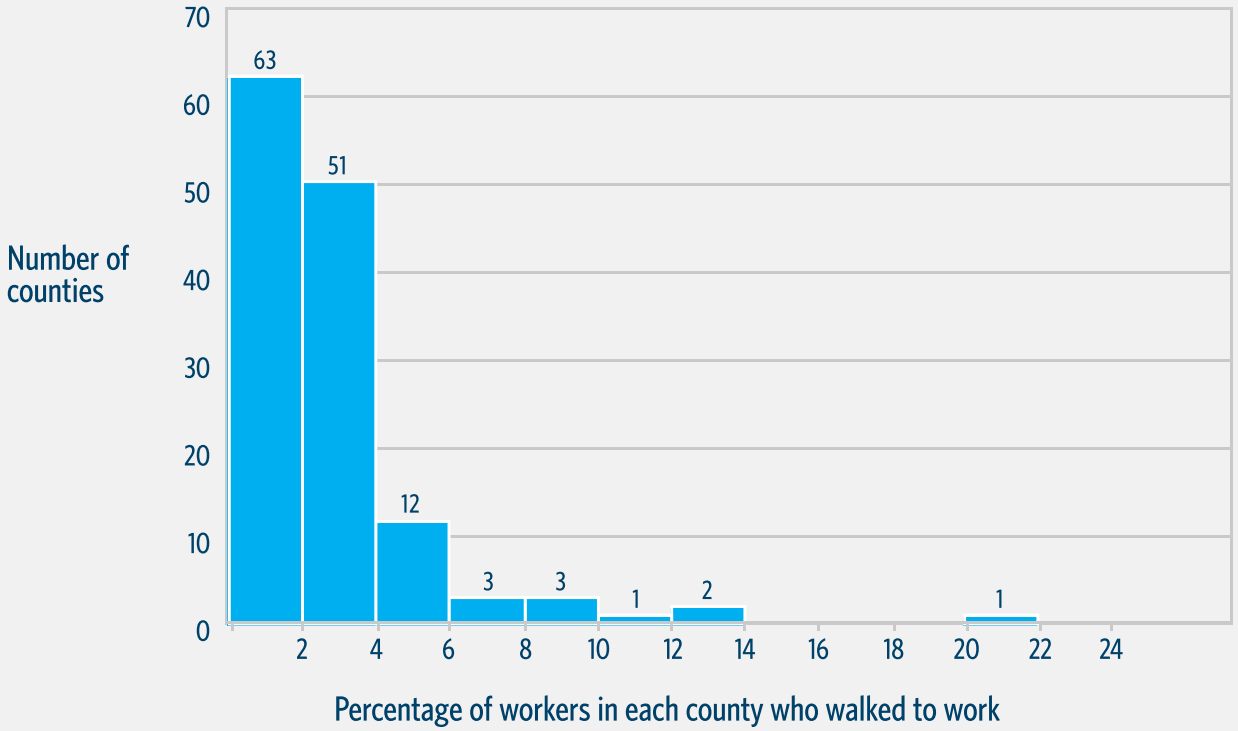
Student Learning Objectives

- I will be able to analyze data and describe data distributions.
- I will be able to understand skewness and how it affects the mean and median.
- I will be able to determine the direction of skewness from a distribution's mean and median.
- I will be able to understand the impact of outliers on the standard deviation and interquartile range.

Statistical measures such as median, mean, range, standard deviation (SD), and interquartile range (IQR) provide useful and interesting snapshots of data sets. In this activity, you will analyze county-level data from the U.S. Census Bureau.

Part 1 – Examine Histograms

1. The following graph shows the percentage of workers (16 and older) who walked to work in 2014 in each of the 136 U.S. counties surveyed. The first bar of the histogram, for example, shows that, in 63 of the counties, fewer than 2 percent of workers walked to work. Refer to this graph to answer the questions that follow.



Source for data: U.S. Census Bureau, 2014 American Community Survey 1-Year Estimates.

data.census.gov/cedsci/table?q=commute%20to%20work&tid=ACSST1Y2014.S0801&hidePreview=true

a. Which characteristics might the counties represented by the bars on the left of the distribution have in common?

Student answers will vary but should note that these counties all have low percentages of commuters who walked to work in 2014, meaning that workers there commuted in other ways. Characteristics could include a more rural setting, where the distances between homes and workplaces are greater and walking to work is more difficult.

b. Which characteristics might the counties represented by the bars on the right of the distribution have in common?

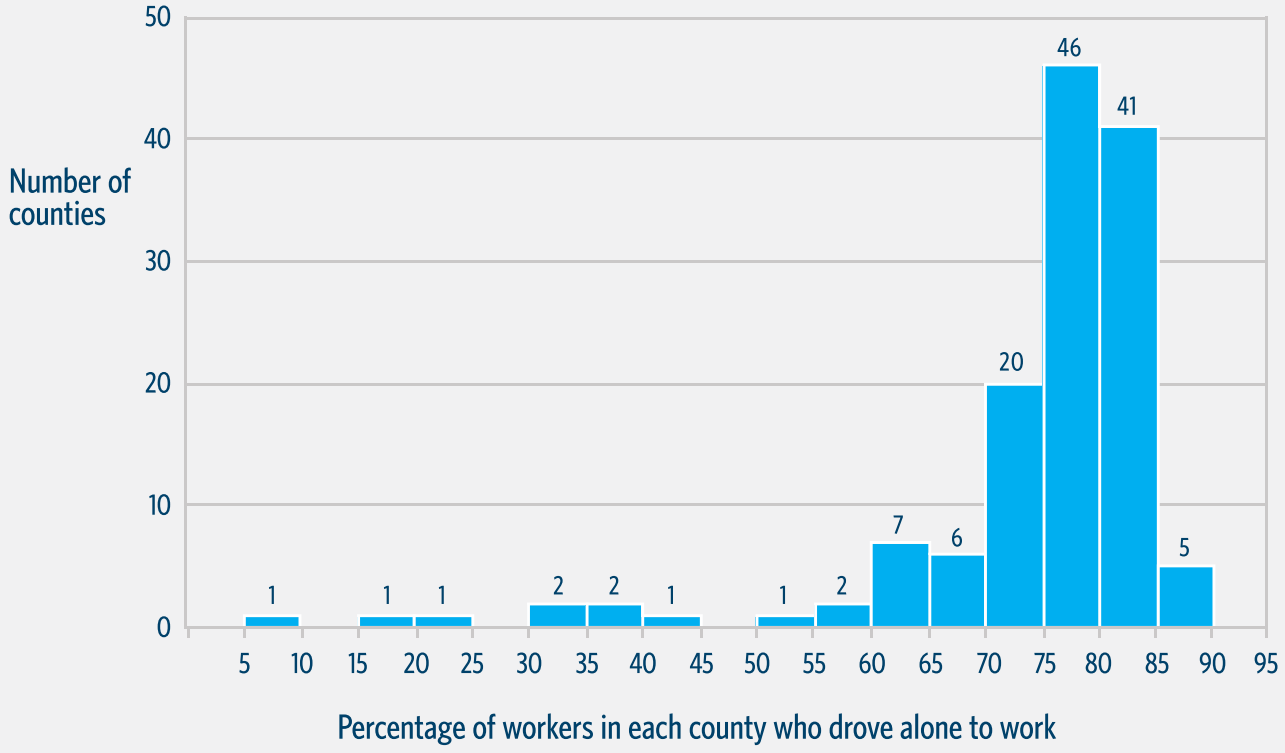
Student answers will vary but should note that these counties all have higher percentages of commuters who walked to work in 2014. Characteristics could include a more urban setting, where the distances between homes and workplaces are small and walking to work is more feasible for more people.

c. Is this histogram skewed left, skewed right, or symmetrical? How do you know? Do you see any outliers?

Skewed right, because the longer "tail" is on the right, where values are larger. There appears to be at least one outlier on the right side of the histogram.

Teachers should explain to students that "symmetrical" does not imply that the histogram is exactly the same on each side.

2. The following histogram shows the percentage of workers (16 and older) in each of the 136 U.S. counties surveyed who drove alone to work in 2014. The bar on the left, for example, indicates that only one county saw between 5 and 10 percent of its workers drive to work by themselves. Refer to this graph to answer the questions that follow.



Source for data: U.S. Census Bureau, 2014 American Community Survey 1-Year Estimates.

data.census.gov/cedsci/table?q=commute%20to%20work&tid=ACSST1Y2014.S0801&hidePreview=true

a. What characteristics might the counties represented by the bars on the distribution's left have in common?

Student answers will vary but should note that these counties all have low percentages of commuters who drove alone to work, meaning that workers might have carpooled or that they did not drive to work at all. Characteristics of such counties could include having an urban center with options like rail, bus, and bike-sharing programs.

Teachers should explain to students that although these bars appear to be outside the distribution's general pattern and later may be identified as outliers, their possible influence on the distribution's shape should not be ignored.

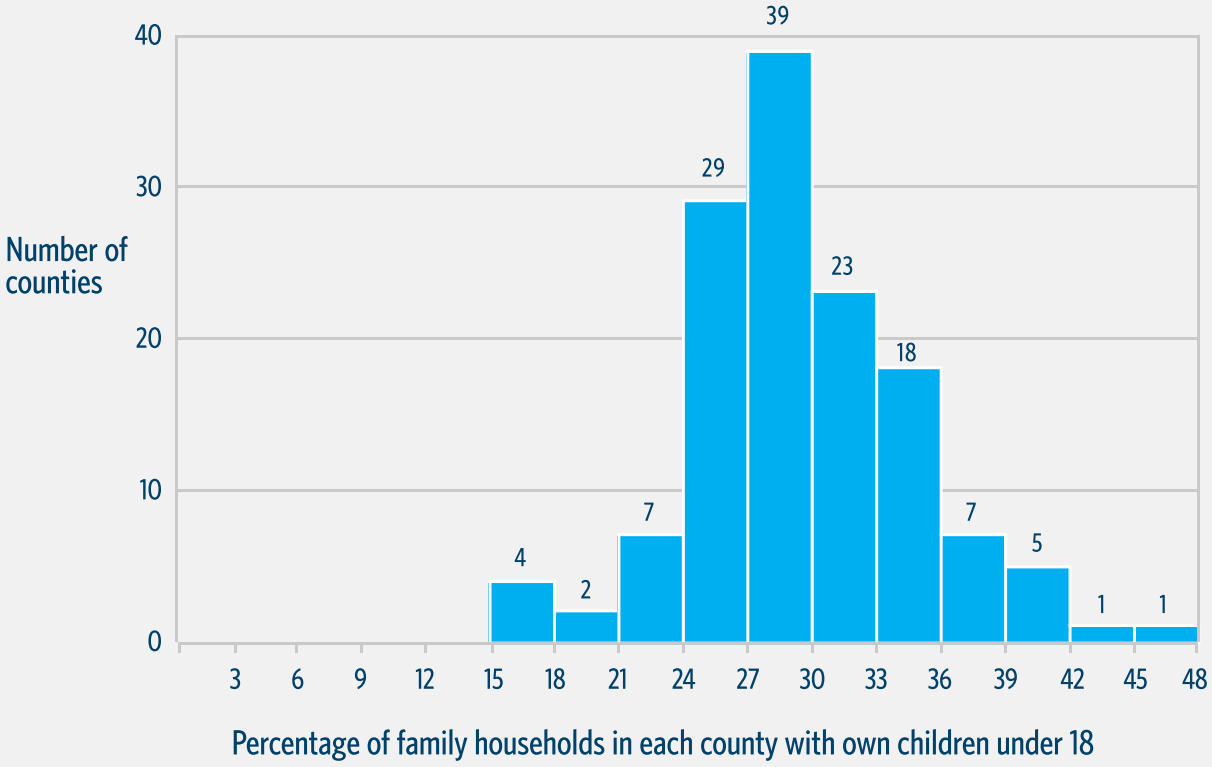
b. What characteristics might the counties represented by the bars on the distribution's right have in common?

Student answers will vary but should note that these counties all have relatively high percentages of commuters who drove alone to work. Characteristics of such counties could include having limited access to mass transportation, such as in more rural places.

c. Is this histogram skewed left, skewed right, or symmetrical? How do you know? Do you see any outliers?

Skewed left, because the longer "tail" is on the left, where values are smaller. The bars on the far left could be outliers.

3. The following histogram shows the percentage of family households with own children under 18 in each of the 136 U.S. counties surveyed. Refer to this graph to answer the questions that follow.



Source for data: U.S. Census Bureau, 2014 American Community Survey 1-Year Estimates.

data.census.gov/cedsci/table?q=Living%20arrangements%20&tid=ACSDP1Y2014.DP02&hidePreview=true&y=2014

a. What characteristics might the counties represented by the bars on the far left have in common?

Student answers will vary but could include that it is possible those counties are urban, with more young, single people and not as many families.

b. What characteristics might the counties represented by the bars on the far right have in common?

Student answers will vary but could include that those counties are possibly more rural or suburban, with residents who are more family-oriented and who prefer more space.

c. Is this histogram skewed left, skewed right, or symmetrical? How do you know? Do you see any outliers?

This histogram is more symmetrical than the previous graphs, though it appears to be slightly skewed right. There do not appear to be any outliers.

Part 2 – Compare Measures of Spread to Predict a Distribution's Shape

In symmetrical distributions, the mean and the median are close together. In skewed distributions, the mean is drawn toward the longer "tail": Skewed-right distributions typically have a mean that is greater than the median, and skewed-left distributions typically have a mean that is less than the median.

Based on this information, answer the following questions, then check your answers by looking back at the histograms in part 1.

1. Question 1 in part 1 focused on the percentage of workers in each county who walked to work in 2014. The mean of those data is 2.86 percent, and the median is 2.00 percent. Based just on the relationship between these two values, would you expect the graph to be symmetrical, skewed right, or skewed left? Why?

   Skewed right, since the mean is greater than the median.

2. Question 2 in part 1 focused on the percentage of workers in each county who drove alone to work in 2014. The mean of those data is 74.03 percent, and the median is 78.20 percent. Based just on the relationship between these two values, would you expect the graph to be symmetrical, skewed right, or skewed left? Why?

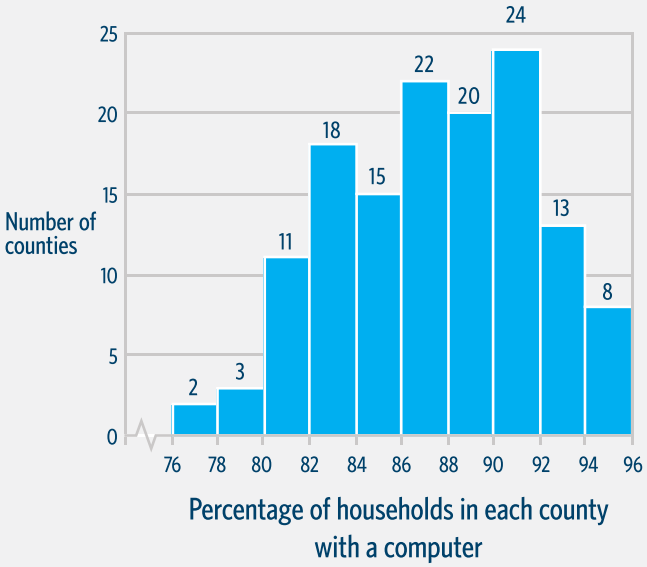
   Skewed left, since the mean is less than the median.

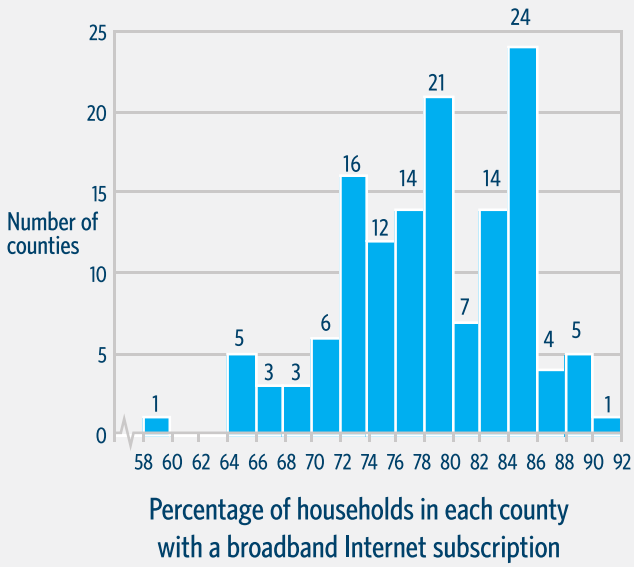
3. Question 3 in part 1 focused on the percentage of family households in each county with own children younger than 18 in 2014. The mean of those data is 29.42 percent, and the median is 28.95 percent. Based just on the relationship between these two values, would you expect the graph to be symmetrical, skewed right, or skewed left? Why?

   Symmetrical, since the values are close together.

Part 3 – Compare Data Distributions

Consider the two histograms below to answer the questions that follow, noting when you make comparisons that these graphs have different data ranges. The histogram on the left shows the percentage of households in each county surveyed that had a computer, and the histogram on the right shows the percentage that had a broadband Internet subscription, both in 2014.





Source for data: U.S. Census Bureau, 2014 American Community Survey 1-Year Estimates.

data.census.gov/cedsci/table?q=Living%20arrangements%20&tid=ACSDP1Y2014.DP02&hidePreview=true&y=2014

1. Which histogram has the greater mean?

   The histogram on the left, for the percentages of households with a computer.

2. Based on the nature of these variables, could you have predicted which data set would have a greater center? Explain your logic.

   Student answers may vary but could include either:

   - **Yes, the expectation could be that there was a larger percentage of households with a computer than the percentage of those with broadband Internet, because an Internet connection typically requires a computer — but not necessarily vice versa.**
   - **No, it might have been expected that a larger percentage of households have broadband Internet in more recent years as opposed to computers, because the prevalence of smartphones, tablets, and other mobile devices means a computer is no longer necessary to access the Internet.**

3. Which distribution is more symmetrical? How do you know?

   The one on the left, as the other graph appears to be somewhat skewed left and to have a low outlier.

4. Which distribution has the larger spread? How do you know?

   The histogram on the right, which has a range of about 34 percentage points (58% - 92%). The graph on the left has a range of about 20 percentage points (76% - 96%).

Part 4 – Evaluate Spread

Answer the following questions using what you know about the three common measures of spread: the range, the IQR, and the SD (formulas of each shown below).

1. Consider again the two histograms about computer and Internet access — with the means and SDs now highlighted in red — and the table below to answer the questions that follow.

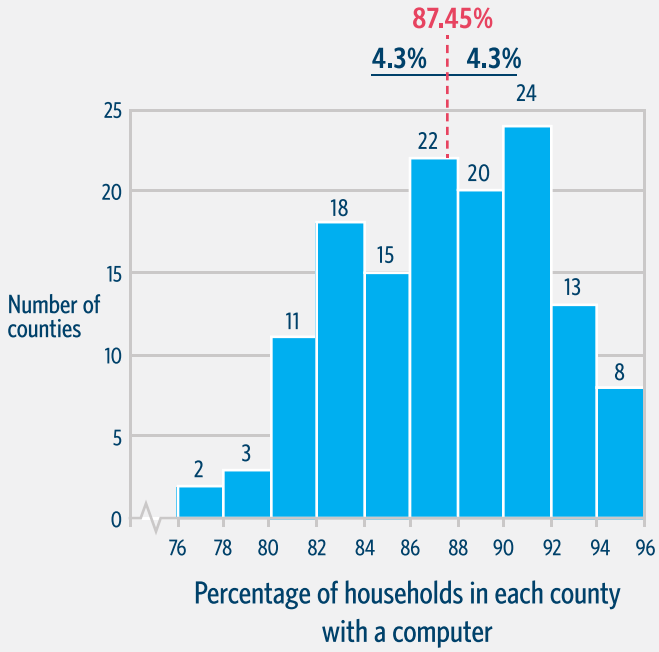
   Range = Maximum - minimum

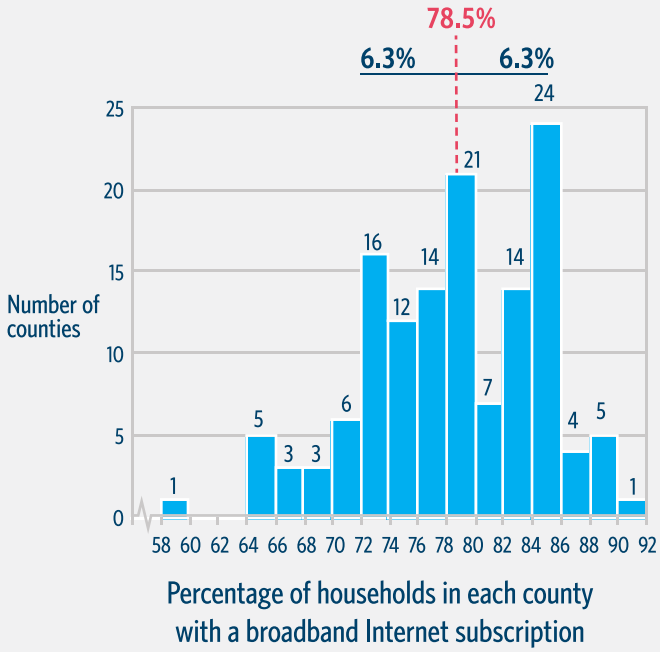
   IQR = Q_3 - Q_1

   Standard Deviation =

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

   Remember: The IQR for these distributions indicates that the widths of the middle 50 percent of the data are 6.7 and 10.1, respectively. The SD says that while some county percentages are closer to and farther from the means, they are within 4.3 and 6.3 percentage points on average, respectively, on both sides of the mean.





Source for data: U.S. Census Bureau, 2014 American Community Survey 1-Year Estimates.

data.census.gov/cedsci/table?q=Living%20arrangements%20&tid=ACSDP1Y2014.DP02&hidePreview=true&y=2014

Values for the "Percentage of households with a computer" graph (in percentage points)			Values for the "Percentage of households with Internet" graph (in percentage points)		
Range	IQR	SD	Range	IQR	SD
18.5	6.7	4.3	32.8	10.1	6.3

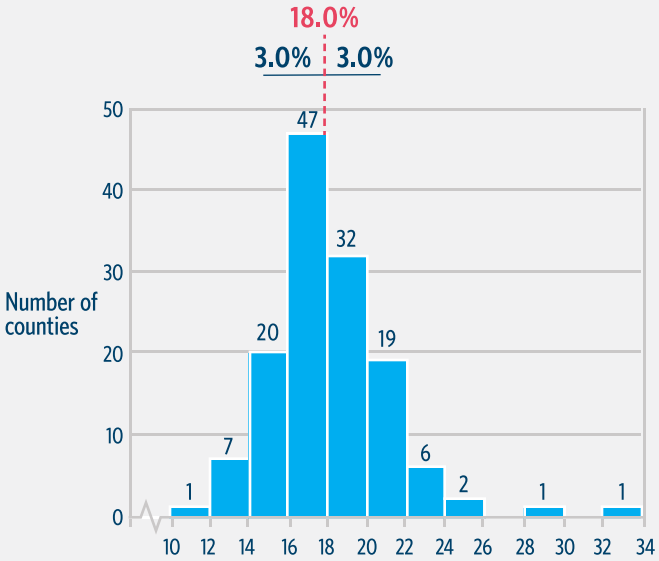
a. What do the values in the table reveal about the spread and variability of the data in both graphs?

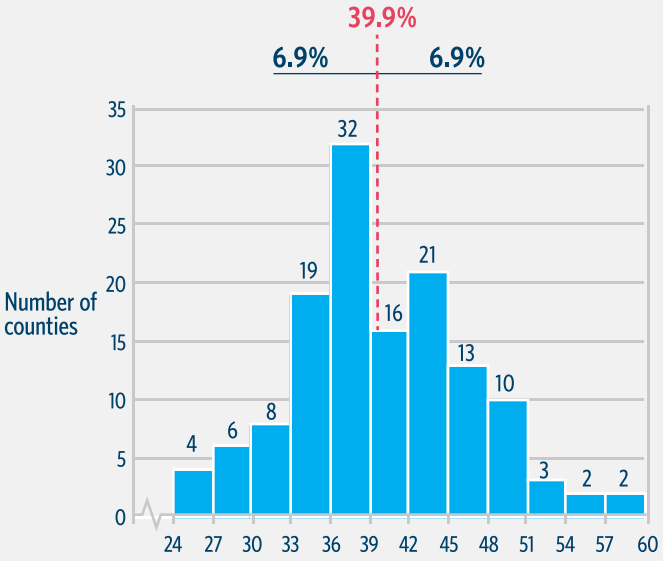
Because the Internet access graph has larger values for the range, IQR, and SD, there is a larger spread and more variability in the data values in that graph than in the other.

b. Mathematically, why do you think that the values for the Internet access distribution are larger than those for the computer ownership distribution?

Because the Internet access graph has a low outlier that increases the value of the range, IQR, and SD.

2. Consider the two histograms and table below to answer the questions that follow. The graph on the left represents the distribution of 136 U.S. counties with percentages of civilians over 16 in service occupations, and the graph on the right shows the percentages in management, business, science, and arts occupations — both in 2014. For example, the tallest bar in the graph on the left indicates that 47 counties had between 16 and 18 percent of their civilians over 16 employed in service occupations.





Source for data: U.S. Census Bureau, 2014 American Community Survey 1-Year Estimates.

data.census.gov/cedsci/table?q=Occupation%20of%20civilian%2016%20and%20over%20&tid=ACSST1Y2018.S2401&t=Occupation&hidePreview=true

Values for the "Percentage of civilians over 16 in service occupations" graph (in percentage points)					Values for the "Percentage of civilians over 16 in management, business, science, and arts occupations" graph (in percentage points)				
Mean	Median	Range	IQR	SD	Mean	Median	Range	IQR	SD
18.0	17.7	21.3	3.2	3.0	39.9	38.6	35.2	8.5	6.9

a. Describe the shape, center, and spread of each histogram.

The service occupations graph is somewhat skewed right. Its center is around 18 percent, with most counties having between about 12 and 22 percent of their civilians over 16 in service occupations. There are two high outliers.

The management, business, science, and arts occupations graph is roughly symmetrical. Its center is around 39, with most counties having between 30 and 51 percent of their civilians over 16 in these occupations. There are no obvious outliers.

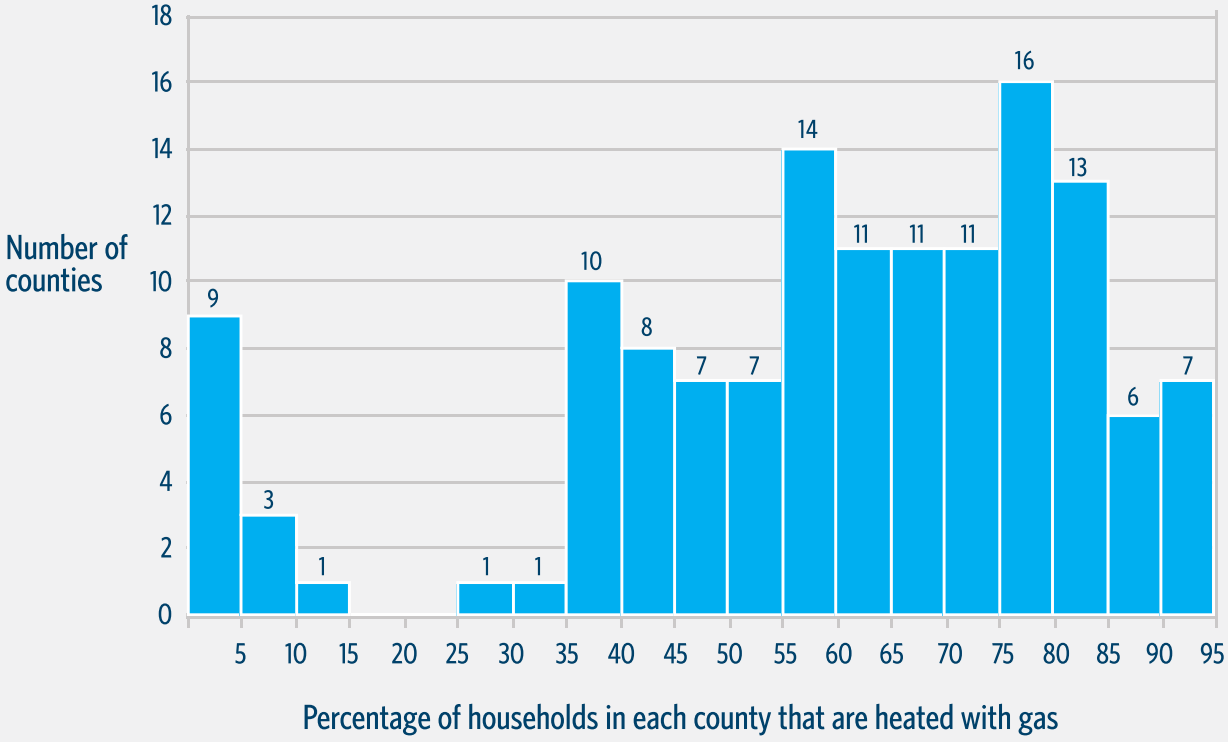
b. Now compare your descriptions for these graphs, thinking about which measures are best for comparing distributions with different scales.

The graph of the management, business, science, and arts occupations data is more symmetrical than the other graph. The center — as both the mean and the median — of the management, business, science, and arts data is about twice as large as that for the services data. The services graph has a smaller spread, shown by an IQR and a SD that are less than half the size of those for the management, business, science, and arts graph.

c. Summarize what your comparisons mean in context.

Overall, there were proportionally fewer workers in service jobs in 2014, and the percentages of people in those jobs were a bit more consistent from county to county compared with the percentages in management, business, science, and arts positions, which varied more by county.

3. Consider the following histogram, which shows the percentage of households in each county that used gas heat in 2014.



Source for data: U.S. Census Bureau, 2014 American Community Survey 1-Year Estimates.

Note: ACS 1-year data include data for all counties in the United States. This activity only focuses on 120 counties to illustrate the statistical concept.

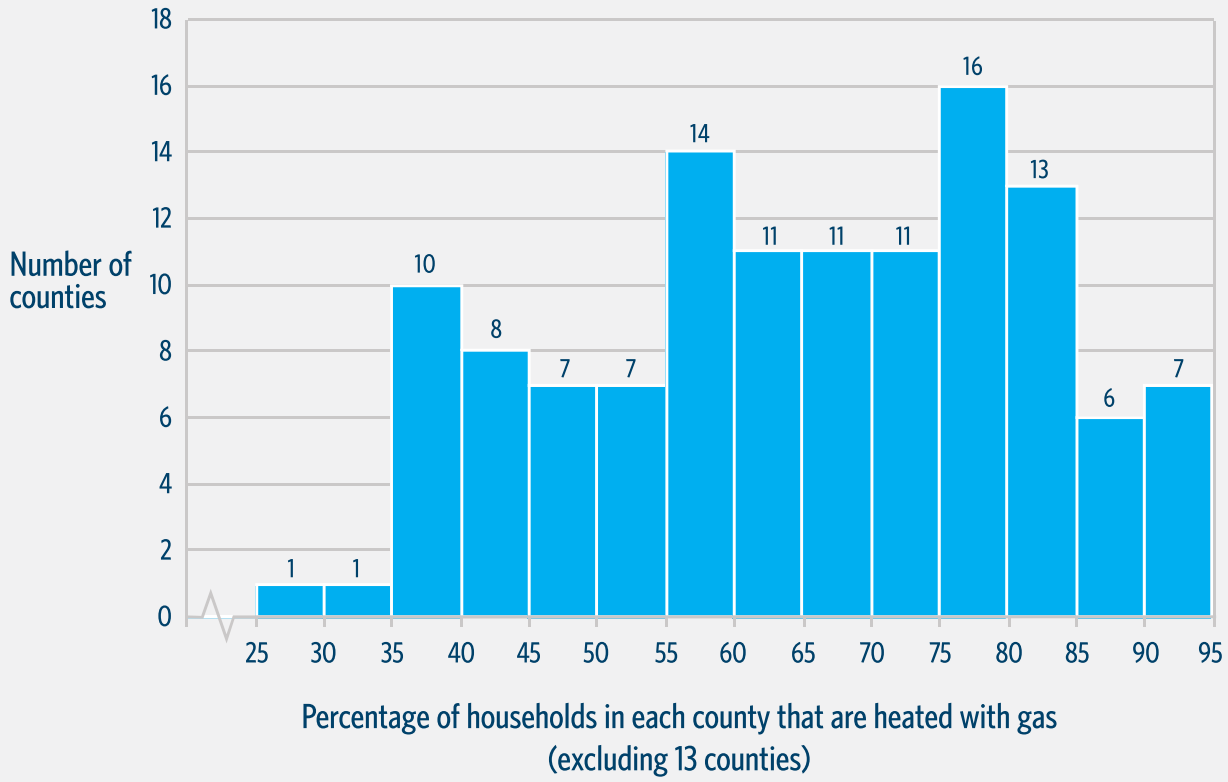
data.census.gov/cedsci/table?q=DP04&tid=ACSDP1Y2014.DP04&hidePreview=true&y=2014

The data are somewhat bimodal, with a group of 13 counties on the left that have low percentages of households heated with gas and a group of more than 120 counties on the right that have over one-third of their households heated with gas. A single measure of center would not be appropriate here.

The 5-number summary, mean, and IQR for the data in this histogram are as follows:

5-Number Summary						
Minimum	Q1	Median	Q3	Maximum	Mean	IQR
2.4%	45.5%	63.5%	77.5%	91.8%	59.4%	32.0%

If the data for the 13 counties with very low percentages (whose mean and median are 4.7 percent and 3.8 percent, respectively) were removed, the graph would appear as follows:



Source for data: U.S. Census Bureau, 2014 American Community Survey 1-Year Estimates.

Note: ACS 1-year data include data for all counties in the United States. This activity focuses on only 120 counties to illustrate the statistical concept.

data.census.gov/cedsci/table?q=DP04&tid=ACSDP1Y2014.DP04&hidePreview=true&y=2014

The 5-number summary, mean, and IQR for this revised histogram would then be:

5-Number Summary						
Minimum	Q1	Median	Q3	Maximum	Mean	IQR
27.6%	52.1%	67.2%	78.5%	91.8%	65.2%	26.4%

a. Which measure of center was more affected by the omission of the 13 counties: the mean or the median? Explain your reasoning, using what you know about the calculations of these two measures.

The mean, because the difference between the original and revised means is 5.8 percentage points, while the difference between the original and revised medians is only 3.7 percentage points. The mean was affected more because this measure of center gets pulled higher or lower based on how the data are distributed or skewed. The median is the middle number and is not as affected by skewed data.

Part 5 – Draw Conclusions About Data

1. Reflect on what you have learned: Write about when it would be best to use the median or the mean to describe the center of a distribution and when it would be best to use the range, the SD, and/or IQR to describe the spread in different kinds of distributions.

Student answers may vary but could include: When outliers are present or if the graph is very skewed, the median is better for describing the center, as the mean is more affected by skewness and outliers than is the median. If a graph is somewhat symmetrical, the mean and the median should be similar, so both can be used to describe the center.

For measures of spread, the range, the SD, and IQR are all good options for describing symmetrical distributions. The IQR is usually the best descriptor of spread for a distribution that is skewed or has outliers, as the range would be affected by outliers and the SD would be affected by both outliers and skewness through the value of the mean.